Grupowanie podziałowe

Przedstawione algorytmy aglomeracyjne działają, poczynając od tworzenia zalążków grup złożonych z najbliżej siebie położonych węzłów (tzw. podejście *bottom up*). Nie uwzględnia się przy tym dalekosiężnych konsekwencji takiego działania, a przecież naszym celem jest dobry podział *całego* grafu. Dlatego, alternatywnie, rozwinęła się druga grupa algorytmów podziałowych. Zgodnie ze swoją nazwą w kolejnych krokach działania dokonują podziału wierzchołków na dwie lub więcej podgrup, zaczynając od całego grafu (podejście *top down*). Formalnie rzecz ujmując, złożoność obliczeniowa algorytmów podziałowych nie jest wielomianowa; przecież najprostszy podział grafu rzędu N na dwie podgrupy jest możliwy na 2^N sposobów – a poszukujemy tego optymalnego. Dlatego w praktyce implementuje się podejścia heurystyczne, np. bardzo rozpowszechniony algorytm Louvain [26].

Algorytm Louvain polega na rekursywnym podziale grafu na dwa podgrafy, tak aby każdy podział skutkował jak najmniejszą wartością *modularności* odzwierciedlającej gęstość połączeń w obrębie powstałych 2 podgrup wobec gęstości połączeń pomiędzy nimi. Naturalnie istota algorytmu tkwi w heurystycznej klasyfikacji wierzchołków do dwóch grup. Przypomina ona do złudzenia… grupowanie aglomeracyjne: początkowo każdy wierzchołek stanowi jednoelementową grupę, a następnie jest łączony z którymś ze swoich sąsiadów. Szczegóły implementacyjne (np. powiększanie grup o pojedyncze, sąsiadujące węzły) skutkują w praktyce niewielką utratą optymalności grupowania oraz bardzo dobrą wydajnością algorytmu. Przede wszystkim jednak jest on w stanie przetworzyć w sensownym czasie nawet bardzo duże sieci w przeciwieństwie do algorytmów aglomeracyjnych, które, przy swojej złożoności obliczeniowej $O(N^2 \log N)$, stają się bezużyteczne już przy sieciach o kilkudziesięciu tysiącach węzłów.

6.4. Rekomendacje

Skoro graf dwudzielny reprezentuje związki pomiędzy obiektami dwóch różnych typów, można postawić pytanie: w jaki sposób te związki się uformowały? Mogą one wynikać z przyczyn zupełnie zewnętrznych, niezwiązanych z naszym grafem, jak np. przynależność dzieci do szkół (wynikająca z rejonizacji) albo sąsiadowanie w tekście pisanym konkretnych rzeczowników i przymiotników (wynikające z gramatyki języka i występujących w nim związków frazeologicznych). Często jednak powstawanie nowych związków wynika w dużej mierze ze związków aktualnie istniejących. Dynamikę takich zjawisk przedstawiliśmy w zarysie w podrozdz. 4.3. Wynikała ona ze stosunkowo prostych modeli decyzji jednostki.

W rzeczywistości ludzki proces decyzyjny jest zdecydowanie bardziej skomplikowany. Nie tylko dlatego, że nasz wybór dotyczy większej rozmaitości przedmiotów, ale przede wszystkim z tego powodu, że często jest oparty na niezwerbalizowanych przesłankach. Nie do końca jesteśmy w stanie wytłumaczyć, dlaczego podoba nam się lokal gastronomiczny, utwór muzyczny czy film – z uwagi na jego obiektywne cechy,

nasze subiektywne postrzeganie…, a może dlatego, że podoba się naszym znajomym? W tym podrozdziale przedstawimy różne algorytmy mające za zadanie modelować tak skomplikowane procesy na miarę różnych dostępnych danych. Łączy je podejście behawioralne, tj. założenie, że każda nowa decyzja ludzka jest zdeterminowana decyzjami poprzednimi tej jednostki oraz innych.

Inaczej niż w przypadku analizy dynamiki sieci celem tych algorytmów jest sugerowanie człowiekowi kolejnych, trafnych decyzji. Dlatego są one znane pod nazwą *systemów rekomendacji* (*recommender system*). Analizując historię zachowań użytkownika, system rekomendacji ma do dyspozycji albo informację binarną (użytkownik i podjął decyzję j), albo informację o ocenie (użytkownik i podjął decyzję j i wiąże się z nią ocena r_{ij}). Obie oceny, niezależnie od skali, mogą być wystawiane *jawnie*, w procesie oceniania, lub *niejawnie*, wskutek obiektywnej analizy aktywności użytkownika. Druga z możliwości odpowiada typowemu „głosowaniu frekwencją", które może dotyczyć wyboru stanowiska kasowego, miejsca do wędkowania czy karmy. W tym kontekście użytkownik nie dzieli się oceną wyboru, bo dotyczy ona błahej sprawy (podróżny), bo nie chce (wędkarz), bo nie może (kot). Coraz częściej mamy jednak do dyspozycji system współdzielenia się opiniami o naszych wyborach, umożliwiający ocenianie tyleż świadome, co subiektywne. Dlatego należy pamiętać, że systemy niejawne dają informacje bardziej rzetelne, natomiast w systemach jawnych większą skłonność do oceniania mają osoby niezadowolone – zatem takie dane są obarczone błędem systematycznym.

Wszystkie systemy rekomendacji skrzętnie zbierają dane historyczne i budują wewnętrzny model zachowań, aby przy najbliższej okazji zasugerować nam decyzję, której nie pożałujemy: tytuł kolejnego utworu do odsłuchania, miejsca do odwiedzenia lub treść reklamy. Spróbujmy zaprojektować prosty system rekomendacji, wykorzystując dane już posiadane. Załóżmy (zgodnie z prawdą), że obsada aktorska jest wynikiem autonomicznej decyzji reżysera. Zadaniem systemu będzie zasugerowanie kolejnego aktora, który mógłby przypaść do gustu reżyserowi. Wyszukamy w tym celu innego reżysera, który angażował podobnych aktorów, i zasugerujemy jednego nowego spośród nich.

Macierz ocen

Potrzebujemy informacji o współpracy reżyserów z aktorami. Posługując się oznaczeniami z podrozdz. 6.2, odtworzymy ją, mnożąc macierze: $\hat{\mathbf{A}}_{(3,1)} = \mathbf{A}_{(3,2)}\mathbf{A}_{(2,1)}$. Dalej będziemy tę macierz nazywać *macierzą ocen* obiektów (*ranking matrix*) dokonanych przez użytkowników i oznaczać przez $\mathbf{R}$. W naszym przykładzie macierz ocen to *de facto* macierz wag w grafie łączącym reżyserów i aktorów. Jest to podstawowa i jedyna informacja, którą się będziemy posługiwać. Element r_{ij} zawiera „ocenę", tj. liczbę filmów, w których aktor j wystąpił u reżysera i. Każdy wiersz $\mathbf{R}$ opisuje w pełni (choć nie chronologicznie) decyzje o angażach aktorskich u konkretnego reżysera. Jej pojedynczy wiersz $\mathbf{u}_i = [r_{i1}, r_{i2}, \dots]$ jest wektorem ocen poszczególnych aktorów dokonanych przez reżysera i.

Podobieństwo δ_{ik} pomiędzy reżyserami i oraz k zdefiniujmy jako iloczyn skalarny $\mathbf{u}_i \cdot \mathbf{u}_k$. Zauważmy, że jest ono tym większe, im więcej tych samych aktorów i tyle samo

razy zostało zatrudnionych przez reżyserów i oraz j. Szukając najlepszej sugestii dla reżysera i, znajdujemy reżysera $\arg\max_k \delta_{ik}$ i sugerujemy jednego z jego aktorów, który dotychczas nie znalazł u niego zatrudnienia. Można powiedzieć, że każemy reżyserowi i wzorować się na wyborach jego „bratniej duszy".

Wyszukiwanie wspólnych ocen

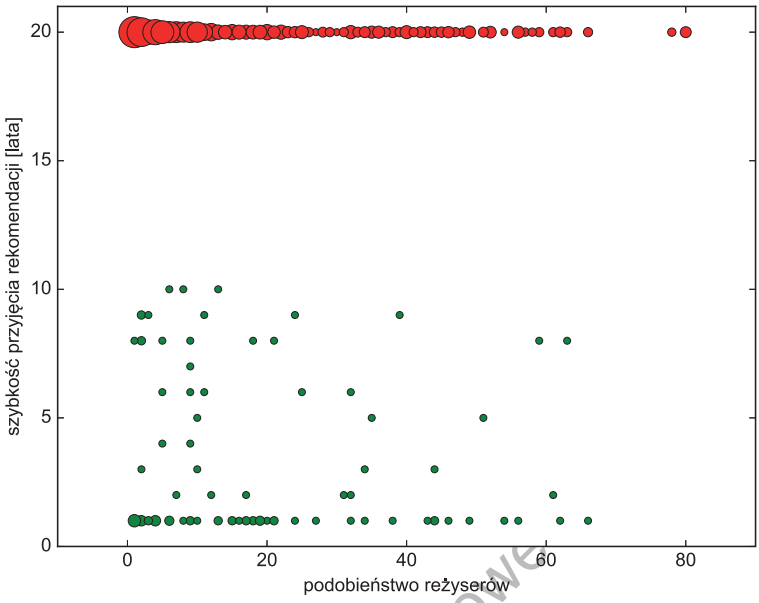
Wykorzystajmy[21] przedstawioną procedurę do zasugerowania reżyserom w każdym roku kalendarzowym jednego aktora na podstawie angaży aktorskich z ostatnich 10 lat. I tak, np. do sugestii w 1981 r. wykorzystujemy historię kina z lat 1972–1981. Dokonajmy też sprawdzenia trafności prognozy, odnotowując, w którym roku reżyser faktycznie zatrudnił rekomendowanego aktora. Tabela 6.2 zawiera przykładową listę trafnych rekomendacji wydanych w 1981 r., wraz z rokiem, w którym reżyser zatrudnił rekomendowanego mu aktora. Można zaobserwować dość duży rozrzut zarówno wartości podobieństwa, jak i czasów do przyjęcia rekomendacji.

Tabela 6.2. Przykładowe rekomendacje aktorów dla reżyserów na podstawie historii współpracy w latach 1972–1981

Rok		Kto poleca	Komu poleca	Kogo poleca	Podobieństwo
polecenia	akceptacji				
1981	1991	Bogusław Linda	Agnieszka Holland	Artur Barciś	6
1981	1982	Andrzej Wajda	Czesław Petelski	Andrzej Łapicki	6
1981	1983	Andrzej Wajda	Andrzej Seweryn	Czesław Wołłejko	61
1981	1982	Andrzej Seweryn	Jerzy Antczak	Andrzej Łapicki	15
1981	1982	Andrzej Wajda	Ewa Petelska	Andrzej Łapicki	6
1981	1989	Andrzej Wajda	Krzysztof Zanussi	Artur Barciś	18

Wygodnie jest zestawić ze sobą obie te wartości na wspólnym wykresie dla wszystkich wydanych rekomendacji, jak to widać na rys. 6.12. Czerwone okręgi na umownej wysokości 20 lat oznaczają rekomendacje odrzucone. Wielkości okręgów odpowiadają liczbie rekomendacji w tym konkretnym miejscu. Z wykresu wynika, że zdecydowana większość (1890 spośród 2020) rekomendacji nie została uwzględniona przez najbliższe 10 lat przez reżyserów. Powodem jest przede wszystkim ograniczona liczbą nowych aktorów, których może przyjąć reżyser w kolejnym roku, związana z jego poziomem aktywności zawodowej. Kolejnymi przyczynami mogą być: (a) nieadekwatność zastosowanej metody rekomendacji, uwzględniającej zbyt krótki horyzont czasowy albo wykorzystującej niewłaściwą miarę podobieństwa reżyserów, oraz (b) niedokładność, niekompletność i niewystarczająca szczegółowość danych historycznych (np. brak zróżnicowania udziału w rolach pierwszoplanowych i drugoplanowych).

[21] Czytelnik znajdzie implementację systemu na stronie WWW towarzyszącej książce.



Rysunek 6.12. Wykres punktowy bliskości źródła rekomendacji oraz czasu do przyjęcia rekomendacji przez reżysera. Pole kół odpowiada liczbie przypadków w danej lokalizacji (najmniejsze koła oznaczają wystąpienie pojedynczej rekomendacji w danym miejscu)

Użytkownicy współoceniający

Zaprezentowany algorytm klasyfikuje się do rodziny metod rekomendacji o angielskim terminie *collaborative filtering*, *CF*. Proponujemy nazywać je metodami *wyszukiwania wspólnych ocen*, gdyż fundamentem ich działania jest założenie, że użytkownicy, którzy podobnie ocenili obiekty w przeszłości, w przyszłości również będą oceniać podobnie. Są więc dla siebie wzajemnie źródłem inspiracji w wyborze nowych obiektów.

W tej obszernej rodzinie możemy dalej wyróżnić metody *wyszukiwania użytkowników współoceniających* (*neighbor-based CF*) oraz *wyszukiwania obiektów współocenionych* (*item-to-item CF*). Nasz przykładowy algorytm należy do tej pierwszej kategorii, gdyż poszukuje użytkowników o podobnym wektorze ocen. Podobieństwo δ_{ik} można definiować rozmaicie, wykorzystując w tym celu:

- Współczynnik korelacji Pearsona: $\delta_{ik} = \text{cov}(\mathbf{u}_i, \mathbf{u}_k)/(\sigma_i \sigma_k)$. Opisuje on współliniowość ocen wystawionych przez użytkowników i oraz k i przyjmuje wartości od jedności dla zupełnie zgodnych ocen, poprzez zero dla ocen kompletnie niezwiązanych ze sobą, aż do minus jeden, gdy użytkownicy oceniają obiekt zupełnie przeciwnie. Naturalnie jesteśmy zainteresowani wyszukaniem par użytkowników o wysoce skorelowanych ocenach, z wyjątkiem par oceniających dokładnie jednakowo (takie „idealne małżeństwa" nie mają sobie nic nowego do powiedzenia). Ponadto, silna negatywna korelacja może być równie cenna jak pozytywna; wybór dokonany przez jednego użytkownika staje się swoistą antyrekomendacją dla drugiego.

- Współczynnik korelacji Spearmana obliczany tak jak korelacja Pearsona, lecz nie dla wartości ocen, ale ich miejsca w indywidualnych rankingach każdego z użytkowników. I tak, najniżej oceniany obiekt (np. najrzadziej zatrudniany aktor) otrzymuje wartość 1. Kolejni, coraz częściej zatrudniani aktorzy otrzymują kolejne, rosnące wartości ocen.
- Podobieństwo kosinusowe, $\delta_{ik} = \frac{\mathbf{u}_i \cdot \mathbf{u}_k}{\|\mathbf{u}_i\| \, \|\mathbf{u}_k\|}$, interpretowane jako kosinus kąta pomiędzy wektorami ocen w przestrzeni o wymiarowości równej liczbie ocenianych obiektów. Zakres i interpretacja wyników są takie same jak poprzednio, ale wartości ocen nie są tutaj relatywizowane według ich średnich, lecz zawsze mierzone względem początku układu kartezjańskiego. Wynika stąd m.in., że współwystępujące oceny „zerowe" nie przyczyniają się do wzrostu podobieństwa.
- Współczynnik Tanimoto, $\delta_{ik} = \frac{\mathbf{u}_i \cdot \mathbf{u}_k}{\|\mathbf{u}_i\| \, \|\mathbf{u}_k\| - \mathbf{u}_i \cdot \mathbf{u}_k}$. Jest on uogólnieniem podobieństwa Jaccarda na wektory atrybutów (u nas: ocen).

System rekomendacji wyznacza sugestie dla użytkownika i, posługując się ocenami innych użytkowników, o największej wartości podobieństwa. Do otoczenia użytkownika, U_i, można zakwalifikować *wszystkich* użytkowników, których podobieństwo jest wyższe niż ustalona *wartość progowa* $\delta^\star$. Alternatywnie, można uwzględnić jedynie k najbardziej podobnych do i sąsiadów (*k-NN, k-nearest neighbors*). Bez względu na wybrany wariant stajemy następnie w obliczu wyboru metody agregacji rankingów użytkowników z otoczenia do postaci konkretnej rekomendacji $\hat{\mathbf{u}}_i$ przedstawianej użytkownikowi i. Można tego dokonać na co najmniej kilka sposobów:

- wyznaczając prostą średniej z ocen, $\hat{\mathbf{u}}_i = \frac{1}{\|U_i\|} \sum_{j \in U_i} u_{ij}$;
- wyznaczając średnią ważoną, $\hat{\mathbf{u}}_i = \frac{1}{\sum w_{ij}} \sum_{j \in U_i} w_{ij} u_{ij}$; waga poszczególnych użytkowników z otoczenia jest z reguły uzależniona od ich wartości podobieństwa;
- wybierając rekomendację w postaci wektora ocen, który ma największa liczba użytkowników w otoczeniu.

W ostatniej fazie użytkownikowi prezentuje się te elementy wektora rekomendacji $\hat{\mathbf{u}}_i$, które najbardziej różnią się od $\mathbf{u}_i$. W praktyce odpowiadają one obiektom, których użytkownik i nigdy nie ocenił. Ich liczba zależy od sposobu prezentacji i jest ograniczona np. ilością miejsca na ekranie lub przyjętą strategią informacyjną (mała liczba silnie rekomendowanych obiektów kontra panel umożliwiający eksplorowanie kolejnych sugestii).

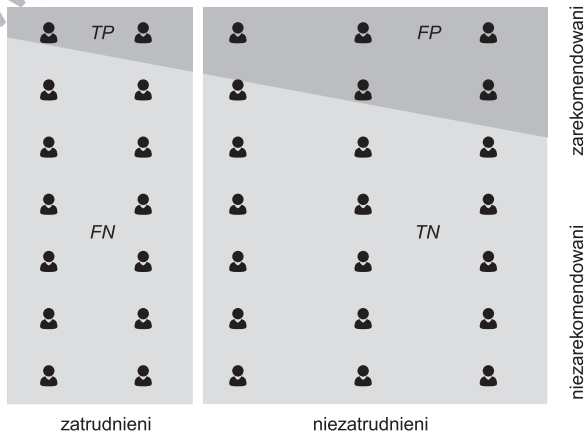
Ocena rekomendacji

Widzimy, że na algorytm rekomendacji składa się kombinacja różnych metod (obliczania podobieństwa, wyboru sąsiadów), a także wartości parametrów tych metod. Ta dowolność sprawia, że projektant systemu porusza się w dość dużej przestrzeni możliwych rozwiązań, często bez adekwatnej wiedzy pozwalającej wstępnie wykluczyć te nieefektywne. Dlatego dość popularnym podejściem staje się sprawdzenie *wszystkich* kombinacji i wybór najlepszej z nich. Wymaga to, rzecz jasna, konkretnej miary oceny jakości działania systemu rekomendacji skonstruowanego w określony sposób. W przykładzie

„aktorskim" ograniczyliśmy się do sprawdzenia, jaka liczba rekomendacji wydanych na podstawie 10-letniej historii angaży została uwzględniona przez reżyserów w ciągu kolejnych 10 lat. Możemy podzielić aktorów, stosując dwa kryteria: zarekomendowanych (i niezarekomendowanych) oraz, niezależnie, faktycznie zatrudnionych w ciągu następnych 10 lat (i niezatrudnionych). Taki przykładowy podział został przedstawiony schematycznie w postaci diagramu na rys. 6.13. Zadanie rekomendacji należy do ogólnej klasy zadań klasyfikacji binarnej, czyli wskazania przynależności obiektów do jednej z dwóch klas nazywanych umownie pozytywną i negatywną. W naszym przykładzie odpowiadają one zbiorom aktorów zarekomendowanych do zatrudnienia i pozostałych. System rekomendacji może pomylić się przy tym na dwa sposoby: zarekomendować aktora, który nie zostanie zaangażowany, oraz nie zarekomendować aktora, który w rzeczywistości zostanie zaangażowany. Odpowiadają temu pojęcia klasyfikacji *fałszywie pozytywnej* (*false positive, FP*) oraz *fałszywie negatywnej* (*false negative, FN*) i związane z nimi dwie podstawowe miary jakości klasyfikacji J:

- precyzji: $J_P = \frac{\|TP\|}{\|TP\|+\|FP\|}$ – prawdopodobieństwa, że rekomendacja zostanie przyjęta;
- skuteczności: $J_R = \frac{\|TP\|}{\|TP\|+\|FN\|}$ – prawdopodobieństwa, że obiekt, który powinien być rekomendowany, został faktycznie zarekomendowany.

Naturalnie, dobry system rekomendacji powinien cechować się małą wartością obu typów błędów. Złe rekomendacje skutkują niezadowoleniem użytkownika, a w ostateczności – spadkiem bardziej namacalnych wskaźników ekonomicznych. Na szczęście nie zawsze musimy rygorystycznie dążyć do maksymalizacji obu wskaźników. Precyzja nie jest najistotniejsza, jeśli nie wymagamy, aby *wszystkie* nasze rekomendacje były trafne. Możemy przecież założyć, że ostatecznie weryfikuje je użytkownik – tak obsługujemy np. wyszukiwarkę internetową. Skuteczność również nie musi być priorytetem, jeśli użytkownikowi wystarczą rekomendacje dobre, chociaż niekompletne. Nie możemy sobie jednak pozwolić na zlekceważenie obu kryteriów i dostarczanie rekomendacji, które nie są ani trafne (niska precyzja), ani wartościowe (niska skuteczność).



Rysunek 6.13. Diagram odpowiadający przykładowej tablicy błędów rekomendacji

Spośród całkiem dużego arsenału metod ewaluacji systemów rekomendacji warto wskazać jeszcze dwie metody wykorzystujące miary błędów, szczególnie przydatne, gdy ocena przedmiotu nie jest wartością binarną. Pierwszą z tych miar jest wartość błędu *średniokwadratowego* rekomendacji (*root mean square error, RMSE*):

$$J_Q = \sqrt{\frac{1}{n} \sum_{i=1,\ldots,n} (\hat{\mathbf{u}}_i - \mathbf{u}_i^\star)^2}, \tag{6.2}$$

gdzie $\mathbf{u}^\star$ jest wektorem faktycznych ocen użytkownika, a drugą – wartość *średniego błędu bezwzględnego* (*mean absolute error, MAE*):

$$J = \frac{1}{n} \sum_{i=1,\ldots,n} |\hat{\mathbf{u}}_i - \mathbf{u}_i^\star|.$$

Wybór konkretnego wskaźnika oceny systemu rekomendacji zależy ostatecznie od celu, w jakim system ów zostanie zastosowany. Emisja nietrafionej reklamy czy sugestia wykupienia wakacji w miejscu nieodpowiadającym preferencjom użytkownika oznaczają z reguły namacalne straty finansowe. Dobór odpowiedniego kryterium oceny nie powinien stanowić problemu tym bardziej, że jest ich do dyspozycji znacznie więcej. Czytelnika odsyłamy w tym celu do literatury szczegółowej, np. monografii [119]. Należy pamiętać, że jakość rekomendacji silnie zależy od konfiguracji elementów składowych systemu rekomendacji (doboru metod i ich parametrów); należy sprawdzić, które z nich skutkują najlepszymi prognozami.

Obiekty współocenione

Podstawową niedogodnością metod rekomendacji ma podstawie wyszukania osób współoceniających jest tzw. problem *zimnego startu* (*cold start*), czyli braku możliwości zarekomendowania czegokolwiek zupełnie nowym użytkownikom. Nie mają oni żadnej historii zachowań lub jest ona wciąż mało specyficzna, co prowadzi do wyznaczenia mało trafnych sugestii. Wady tej nie mają metody rekomendacji wyszukujące obiekty współocenione: dokonują one projekcji wierzchołków-użytkowników na wierzchołki-obiekty, budując graf zależności pomiędzy tymi ostatnimi na podstawie dotychczasowej aktywności wszystkich użytkowników. Można więc zarekomendować zupełnie nowemu użytkownikowi obiekty podobne do pierwszego ocenionego przezeń obiektu.

Za podstawową miarę podobieństwa międzyobiektowego jest uznawane podobieństwo kosinusowe dwóch punktów reprezentujących dwa porównywane obiekty i oraz j. Lokalizacje tych punktów określają wektory kolumnowe ocen obiektów: $\mathbf{o}_i = [r_{1i}, r_{2i}, \ldots]$ dla i oraz $\mathbf{o}_j$ dla j. Podobieństwo to jest wyznaczane w przestrzeni, w której każdy wymiar odpowiada decyzji poszczególnego użytkownika. Tak jak w przypadku podobieństwa użytkowników współoceniających, wyznaczone podobieństwo jest tym mniejsze, im więcej osób wybrało jeden z obiektów, nie wybierając drugiego. Ostatecznie użytkownikowi są rekomendowane obiekty najpodobniejsze do tych, które sam wybrał.